(2)

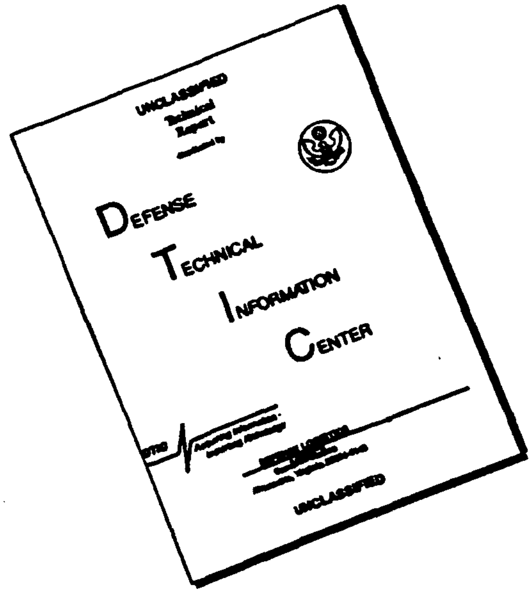
Epistemological Relevance and Statistical Knowledge*

by

Henry E. Kyburg, Jr.
University of Rochester

* Research upon which this work is based was in part supported by the U. S. Army Signals Wareface Center.

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*

2. REPORT DATE
1989

3. REPORT TYPE AND DATES COVERED
Unknown

4. TITLE AND SUBTITLE
Epistemological Relevance and Statistical Knowledge

5. FUNDING NUMBERS
DAAB10-86-C-0567

6. AUTHOR(S)
Henry E. Kyburg, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Rochester
Deparment of Philosophy
Rochester, NY 14627

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army CECOM Signals Warfare Directorate
Vint Hill Farms Station
Warrenton, VA 22186-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
92-TRF-0016

1. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Statement A; Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

For many years, at least since McCarthy and Hayes (1969), writers have lamented, and attempted to compensate for, the alleged fact that we often do not have adequate statistical knowledge for governing the uncertainty of belief, for making uncertain inferences, and the like. It is hardly ever spelled out what "adequate statistical knowledge" would be, if we had it, and how adequate statistical knlwledge could be used to control and regulate epistemic uncertainty. Our purpose here is not to evaluate alternative treatments of uncertainty, but rather to explore the question of how far you can go on the basis of statistical knowledge that you do have, and what considerations must be taken account of in this attempt. Relatively few people have explored the question of how far you can go using statistical knowledge.

1. SUBJECT TERMS
Artificial Intelligence, Data Fuison Epistemological Relevance Statistical Knowledge

15. NUMBER OF PAGES
12

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to **stay within the lines to meet optical scanning requirements.**

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - DOD - Leave blank.
DOE - DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - NASA - Leave blank.
NTIS - NTIS - Leave blank.

Block 13. Abstract. Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (NTIS only).

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

Epistemological Relevance and Statistical Knowledge

1. Background.

For many years, at least since McCarthy and Hayes (1969), writers have lamented, and attempted to compensate for, the alleged fact that we often do not have adequate statistical knowledge for governing the uncertainty of belief, for making uncertain inferences, and the like. It is hardly ever spelled out what "adequate statistical knowledge" would be, if we had it, and how adequate statistical knowledge could be used to control and regulate epistemic uncertainty.

One response to the lack of adequate statistics has been to search for non-statistical measures of uncertainty. The minimal variant has been to propose "subjective probability" as a concept to which we can turn when we lack statistics.

This proposal comes in widely differing flavors, based on the dreadful ambiguity of "subjective". Sometimes this means merely "indexed by a subject". In this sense there is no conflict with statistical representations: the "subjectivity" involved just represents the fact that statistical knowledge is related to a knower. (This appears to be the sense of "subjective" employed by Cheeseman (1987).)

At the other extreme, "subjective" may mean "arbitrary," "whimsical," "subject to no objective control or constrtaint." Those who think we must turn in this direction are influenced by the feeling that in many cases there may be nothing better to turn to.

Other proposals concern non-probabilistic measures of uncertainty: the certainty factors of Mycin (1984), the belief functions of Shafer (1987), the fuzzy membership relation of Zadeh (1986).

Our purpose here is not to evaluate these alternative treatments of uncertainty, but rather to explore the question of how far you can go on the basis of statistical knowledge that you do have, and what considerations must be taken account of in this attempt. Relatively few people have explored the question of how far you can go using statistical knowledge. One writer who has taken this question seriously is Bacchus (1988).

A second question, in fact the one that McCarthy and Hayes had in mind, is the question of using statistical knowledge to provide an underpinning for uncertain inference -- that is,

inference that is based on incomplete knowledge. A basic presupposition of the non-monotonic and default industries seems to be that you cannot very often base such inferences on statistical knowledge. In part this presupposition is based on the feeling that "typicality" and "frequency" mean different things. Be that as it may, in formalizing non-monotonic logics many people seem to be led to considerations that (not surprisingly!) mirror considerations appropriate to the application of statistical knowledge.

Thus Etherington (1987) introduces the concept of preference among models; Konolige (1987) defines a notion of minimal extension; Touretzky (1986) gives a metric for inferential distance.

These metrics will be reflected in one of three principles governing the relvance of statistical knowledge to be discussed below. Our analysis of the ground-rules for the use of statistical knowledge will throw light on the "cancellation principles" of non-monotonic logic as well.

2. Assumptions.

The assumptions we make here are relatively few. We suppose that the knowledge base may have statistical knowledge in it. This statistical knowledge may be construed in a number of ways -- for example as statements concerning chances or statements concerning frequencies in an arbitrarily long run, or statements concerning frequencies summed over possible worlds. We do suppose that these statements are general: that is that they do not represent the fact that we have recorded a frequency in a specific sample. We may have done so; but we also may have gotten our statistical knowledge from a handbook, or a dependable colleague. In any event, the statistical knowledge in our knowledge base is taken to be general scientific knowledge relating properties; notationally, we will write "$\%(R, A) = p$" to represent the fact that the long-run frequency A's among R's is p.

Our second assumption is that statements fall into equivalence classes with respect to the statistical information that is epistemically most relevant to them. It is sometimes said that we can't use statistical knowledge to determine the chance of heads on the toss of a coin that is to be tossed once and then destroyed, on the ground that we have no statistics about that

coin

But the next toss of this coin is also the next toss of a U.S. coin, the next toss of a modern coin, etc. There are many equivalent ways of phrasing "the next toss of this coin will land heads" that will lead to different reference classes. Similarly, if we know that this sample of B's exhibits a relative frequency of C's of .45, then the chance that the long run frequency is close to the frequency in the sample is the same as the chance that the long run frequency is close to .45.

We can express this as a formal principle: If "$S \rightleftarrows T$" is in our knowledge base, then the same statistical knowledge is potentially relevant to S and to T.

Our third assumption is the general one that our knowledge base can be expressed in a first order extensional language. (Of course this requires the inclusion of enough first order set theory to accommodate the statistics!) We take an individual, however, to be arbitrarily complex: for example it might be a trial of a complicated compound experiment.

Finally, in order for statistics to be of interest, we suppose that we may know some things about an individual without knowing everything about it. Thus we might know of "the next trial" that it is a trial consisting of selecting one of a number of urns at random, and then selecting a coin at random from the urn, and then tossing the coin 10 times. And then we might be interested in whether the tenth toss landed heads on that trial, or we might know the distribution of heads in the tosses, and we might be interested in whether the urn was urn number 4, or we might be interested in knowing something about the frequency of two headed coins in the urn from which we got our sample.

3. Interference I.

We will be concerned with the way in which some items of statistical knowledge can interfere with the epistemic relevance of other items. The simplest and clearest cases can be borrowed from non-monotonic logic.

If all we know of Tweety is that she is a bird, it is reasonable to believe that she can fly. If we also know that she is a penguin, then it is reasonable to believe that she cannot fly, since our knowledge about the chances of a penguin flying interferes with our knowledge about the chances of a bird flying.

If (as we may in our biological ignorance suppose) there is

a rare kind of penguin that <u>can</u> fly, and if we know that Tweety is one of them, then this new knowledge interfers with our general knowledge about penguins, and again we may suppose that Tweety <u>can</u> fly.

This relation has been noted by Etherington, Poole, Konolige, and others. It corresponds to what Reichenbach (1949) had in mind when he said that we should base our posits (degrees of belief) on the "narrowest" reference class concerning which we have adequate statistics. ("Having adequate statistics" does <u>not</u> mean having knowledge of a sample of the class in question; it means having useful <u>general</u> knowledge about that class, whatever it may be based on.)

A principle embodying this natural constraint must be stated with somewhat more generality than is at first obvious, however.

Suppose (to move to an artificial example) we know of a certain ball that it is a ball in certain room, and that we know that fifty percent of the balls in that room are black. (A natural way of designating it is by means of the definite description "the ball next to be chosen." We could also take the individual concerned to be a choosing of a ball; the latter would be natural if we were to consider repeated samplings from the room.) Suppose we know also that that particular ball is also one in an urn, in which 75% are black. The second piece of statistical knowledge is clearly epistemically relevant and the first is not. This intuition is based on the fact that the set of balls in the urn in the room is a <u>subset</u> of the set of balls in the room.

But how about the one-membered subsets of the power set of balls in the room? They are abstract objects, and so can't be black, but they can have the property of having a black member. And the ball in question is not a member of this possible reference set, but of course its unit set is. And its unit set will have a black member if and only if the ball in question is black. So what? Well, the set of balls in the urn is not (cannot be) a subset of <u>the set of one membered subsets of the set of balls in the room</u>.

We <u>could</u> stipulate that all the sentences in question have some specific canonical form; but we shall see shortly that that is not such a good idea. What we can do instead is to formulate our principle a bit more broadly.

The Subset Principle. Suppose that "a is a B" is in our knowledge base, and that "%(B, C) = p" is in our knowledge base. Suppose that we know that a' is a C' if and only if a is a C, that a' is a B', and that %(B', C') = p', where $p \neq p'$. This statistical knowledge is *epistemically irrelevant* if we know of a subset of B', B'', such that we know both a' is a B'' and %(B'', C') = p.

The subset principle is one that has been frequently identified in the context of non-monotonic logic.

4. Interference II.

Here is an example that calls for a second principle: As before, suppose we have a roomful of urns, and that a designates a ball in the room. Suppose we know that there are 100 balls in the room, and that 50 are black. But suppose we also know that there are 10 urns, that 9 of them containing four black balls and one white ball, and that the tenth contains the remainder of the balls. The relative frequency of black balls in the first nine urns is .8, and the relative frequency of black balls in the tenth urn is 14/45 = .311...

Let us consider what statistics are relevant to the statement, "a is black." If we know of a only that it is a ball in the room, it is only the statistics about the frequency of black balls in the room that are relevant. If we know also something about how a came to be the designated ball, the other statistics may also be relevant. For example, we might know that a is the ball resulting from first choosing an urn at random, and then choosing a ball at random from the chosen urn. If that is the case, the relevant statistics are those governing the proportion of pairs consisting of an urn, and a ball drawn from that urn, such that the second member of the pair is black. We can easily calculate the proportion of pairs having this property to be .9 * .8 + .1 * .31... = .751...

But why, under these circumstances, should we regard the statistics concerning balls in the room to be epistemically irrelevant? The interfering set isn't a subset of its competitor. (Note that .751... cannot, mathematically, be the relative frequency in any subset of the st of balls in the room!)

But we can find a relationship: there is a possible reference class that matches the competitor, of which the correct

reference set *is* a subset -- namely, the cross product of the set of urns and the set of balls. This construction is particularly important in the context of (so-called) Bayesian inference; the model we just looked at corresponds to a non-sampling case in which we have a prior probability of .9 combined with a conditional probability of .8, and a prior probability of .1 combined with a conditional probability of .311... We therefore call the rule the Bayesian Principle:

<u>The Bayesian Principle:</u> Suppose that " $<a, b>$ is a B " is in our knowledge base, and that " $\%(B, C) = p$ " is in our knowledge base. Suppose that we know that a' is a C' if and only if a is a C, that a' is a B', and that $\%(B', C') = p' \neq p$. This statistical knowledge is *epistemically irrelevant* if we know of a cross product of B' with B'' and a corresponding subset C'' and a'' such that

(1) $<a', a''>$ is known to be in $B' \times B''$,
(2) $<a', a''>$ is in C'' if and only if a is in C,
(3) $\%(B' \times B'', C'') = p'$,

and for some B^* known to be a subset of $B' \times B''$,

(4) $\%(B^*, C) = p$.

To see how this works in our illustrative example, let U be the set of urns, B the set of balls, E the set of pairs corresponding to the experimental set-up, with $<x, y>$ in E just in case x is an urn and y is a ball in <u>that</u> urn. Our target property is the set of pairs C in which the second member is black. The proportion is just what we calculated before: .751...

To show that a', and the statistical knowledge that the proportion of balls in B that are black is .50, is *not* epistemically relevant, we observe that $<a', u>$ is known to be a member of $B \times U$, where u is the (unknown) urn selected, that the proportion of $B \times U$ in C is 0.50, but that there is a subset of $B \times U$ -- namely E itself -- in which the proportion of members of C is the same as that in E.

The Bayesian principle is followed in constructing representations of uncertainty, particularly in cases in which uncertainties are modified by new evidence, but I have not noticed it in discussions of non-monotonic inference. It should be, of course.

Almost all (species of) mammals give birth to their young

live. Given an arbitrary individual mammal, we do not have reason to think it gives birth to its young live, since half of mammals don't give birth at all. (The analogous point with respect to birds was pointed out by Nutter (1987).) Given an arbitrary individual female mammal, we have reason to think it will give birth to its young live, since almost all species of mammals are such that when their females reproduce, they do it that way. (Almost all the reproductive balls in almost all the urns are white, though it is not the case that almost all the balls in an urn are reproductive.) We accommodate the ovoviviparous platypus by noting that its species (its urn) is unusual.

5. Interference III.

The final principle of relevance we need for dealing with statistical knowledge is in a sense the dual of our first principle, the subset principle. Suppose that you are sampling from a population P with a view to making an inference about the proportion of B's there are in P. It is a general set theoretical fact that we will not explore more deeply that almost all subsets of a given set reflect within narrow limits the composition of the parent set.

Putting flesh on this observation, we might note that (using a crude approximation), whatever the proportion of P's that are B's, the proportion of 10,000 member subsets of P that have a proportion of B's within .04 of the actual proportion is at least .975.

Suppose you look at 10,000 P's and find that 5000 of them are B's. Quite clearly, at a level of confidence of .975, one ought to suppose that between .46 and .54 of the P's are B's.

Of course we might have various bits of knowledge that are relevant to this fact that fall under the first two categories. For example, it may be that we know that our sample is not a random one, because we know that it was drawn in a special way that produces representative samples only rarely. Or we may know that we are sampling from a collection of populations in which we know something about the distribution of the relative frequency of B's.

But let us assume that neither of these are the case -- that is, that neither the subset principle nor the Bayesian principle apply. So we may say that the chances are at least

.975 that the proportion of B's is between .46 and .54.

But now note: We also have observed a subset of 5000, of which 100% were B's. So why do we not infer (by a precisely parallel argument) that the chances are nearly 0.0 that the proportion of B's is between .46 and .54? It is because the larger sample is epistemically relevant relative to the smaller one, while the smaller sample is not relevant relative to the larger one. A principle that captures this intuition is:

The Supersample Principle: Suppose that we know that a_n is a member of P^n and that we are interested in the chance that a_n is Q (e.g., "representative within ϵ"). Suppose that a is known to be a member of R, that a is a Q' if and only if a_n is Q, and that $\%(P^n, Q) = p \neq p' = \%(R, Q')$ are all known. Then our statistical knowledge about R is *epistemically irrelevant* if there is a parallel structure to our original one that is such that we also know that a is a subset of a_n.

It is my belief that these three principles are all the principles we need to determine the epistemic relevance of statistical knowledge in the case in which we either have exact knowledge or none at all.

6. Inexact Knowledge.

By providing a new characterization of "difference" among statistical statements, we can easily generalize the above considerations to the general situation. Let two statistical statements "$\%(A, B) \in [p, q]$" *differ* from "$\%(C, D) \in [r, s]$" just in case neither of $[p, q]$ nor $[r, s]$ is included in the other. Then we shall say that one item of statistical knowledge is irrelevant to another if

(a) it differs, but is rendered irrelevant by one of the three principles expounded above, or

(b) it is less exact than the other.

Note that a consequence of thus liberalizing the notion of statistical knowledge is that we now aalways have statistical knowledge, even if it is only of the form, "$\%(A, B) \in [0,1]$". In general, (b) leads us away from statistical knowledge of this form to more substantive statistical knowledge.

7. Computation.

The object of providing such explicit characterizations of relevance and irrelevance is to be able to provide a feasible algorithm for computing the relevant reference class under specific epistemic conditions -- i.e., with given (plausible) background knowledge. Since (as is obvious) potential reference classes can proliferate indefinitely, we have not achieved that point yet; and providing an algorithm is not part of the present project. Nevertheless, it should be clear where we can go from here. Further details are provided in Kyburg (1983) and Loui (1986).

As an illustration of the mechanism we can employ, we can consider the following construction. Let an *inference structure* for a statement S relative to a body of knowledge be a quintuple $<a, B, C, p, q>$, where "a is a C" is known to be equivalent to S, the statement whose epistemic status interests us, "a is a B" is known, and "$\%(B, C) \in [p, q]$" represents the strongest information we have about B and C.

Consider the set I of all inference structures for S. This set is non-empty, since $<a, \{a\}, C, 0, 1>$ is a member of it, whatever else we may know. We perform pass number one: if an inference structure *differs* from an earlier inference structure (i.e., neither [p,q] nor [p',q'] is included in the other) then delete the irrelevant inference structure, if any; otherwise delete both. The result is a set of inference structures that do not differ from one another. They can be partially ordered by inclusion, where we say that one inference structure <a, B, C, p, q> is included in another <a', B', C', p', q'> when [p,q] is a subinterval of [p',q'].

We then perform a second pass, reflecting our concern for information, by deleting any inference structure that properly includes another inference structure in the sequence. The result is a set of inference structures -- it may well contain more than one, and, according to the details of our procedure, may contain an infinite number -- that are all equally strong. This determines both: the epistemic probability of the statement in question; and its <u>inductive</u> acceptability -- i.e., acceptability for non-monotonic logic -- as reflected by the lower bound of its epistemic probability.

8. Conclusions: We arrive at several conclusions

(1) If we accept the equivalence condition -- that statements connected in our knowledge base by a biconditional should have the same probability -- then many more statements than might at first have been thought can have probabilities based on statistical background knowledge.

(2) This has a profound bearing on the representation of uncertainty in our bodies of knowledge. If we suppose that "subjective" confidence in the strong sense of "subjective" is acceptable as a measure of uncertainty only when statistical information is not available, then there are far fewer situations in which purely subjective uncertainties are called for than some people have suggested.

(3) Given the equivalence condition, there may be many potential reference sets for a given equivalence class of statements. We therefore need a way of adjudicating our choice among these reference sets.

(4) There are three ways in which conflict between two potential reference classes can be resolved to the benefit of one of them. Only one of these ways seems to have worked its way into the literature on non-monotonic logic. All three should be taken account of.

(5) These three resolutions reflect the three principles: <u>the Subset Principle</u>, <u>the Bayesian Principle</u>, and <u>the Superset Principle</u>. (In fact the subset principle is reducible to the Bayesian principle (see Kyburg 1983).)

(6) The results of this analysis can be used to implement probabilistic non-monotonic acceptance as well as to determine rationally allowable distributions of uncertainty.

Henry E. Kyburg, Jr. University of Rochester

Bibiography:

Bacchus, Fahiem (1987): *Statistically Founded Degrees of Belief*, Technical report 87-02, Alberta Center for Machine Intelligence, University of Alberta, Edmonton, Alberta, Canada.

Buchan, B.G., and Shortliffe, E. H, (eds): *Rule-Based Expert Systems*, Addison-Wesley, Reading, MA, 1984.

Cheeseman, Peter (1985): "In Defense of Probability," *IJCAI 1985*, Morgan Kaufmann, Los Altos, 1002-1009.

Etherington, D. W.: "A Semantics for Default Logic," *IJCAI 87*, Morgan Kaufman, Los Altos, 495-498.

Konolige, K.: "On the Relation between Default Theories and Autoepistemic Logic," *IJCAI 87*, Morgan Kaufmann, Los Altos, 394-401.

Kyburg, Henry E. Jr. (1974): *The Logical Foundations of Statistical Inference*, Reidel, Dordrecht.

Kyburg, Henry E. Jr. (1983): "The Reference Class," *Philosophy of Science* **50**, 374-397.

Loui, Ronald P. (1986): "Computing Reference Classes," *Uncertainty in Artificial Intelligence, Workshop*, Philadelphia, 1986.

McCarthy, John, and Hayes, Pat (1969): "Some Philosophical Problems from the Standpoint of Artificial Intelligence, " Melzer and Michie (eds) *Machine Intelligence* **4**, Edinburgh University Press, 463-502.

Nutter, Terry Jane (1987): "Uncertainty and Probability," *IJCAI 87*, Morgan Kaufmann, Lost Altos, 373-379.

Shafer, Glen (1976): *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.

Touretzky, D., et. al. (1987): "A Clash of Intuitions: the Current State of Nonmonotonic Multiple Inheritance Systems," IJCAI 87, Morgan Kaufmann, Los Altos, 476-482.